Effective Use of the Lou Mass Storage Cluster

Dec 11, 2013

NASA Advanced Supercomputing Division

Backing up your Data

- Don't think in terms of only backup/rm. They are independent functions
  - Back up the data when it has value and you want to keep it.
  - Files written to Lou filesystems are automatically written to tape
  - Lou filesystems are normally backed up twice daily
  - When you want to delete it from nobackup, verify it is saved to Lou and then delete it.

- We have lost entire nobackup filesystems
  - Each time multiple people say “I just lost six months work”.
  - Your workflow shouldn’t allow this.

Question? Use the Webex chat facility to ask the Host

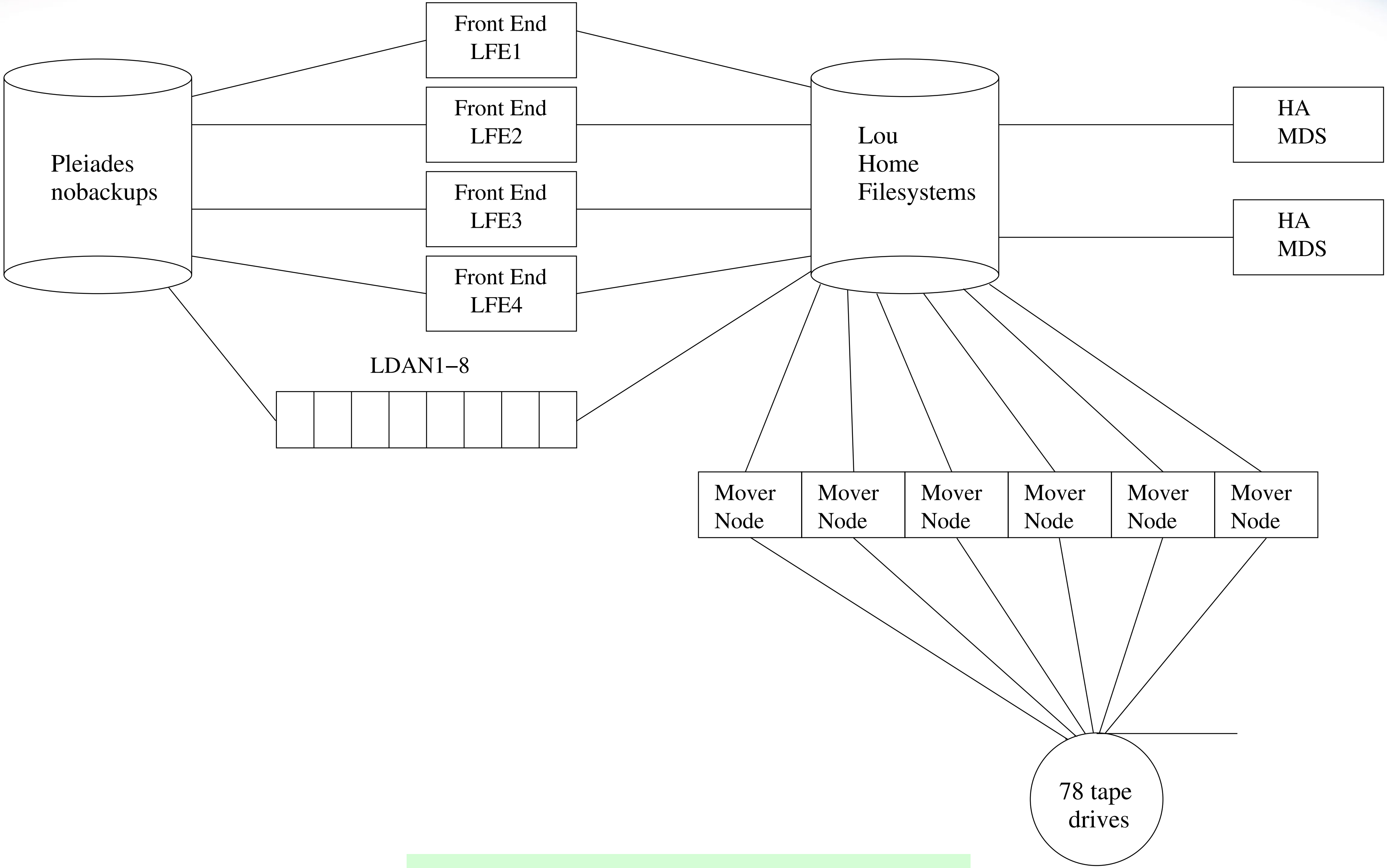
Documentation

- Best Practices are evolving
  - URLs and search strings are included for the HECC KnowledgeBase at http://www.nas.nasa.gov/hecc/
- More than one way to do things
  - Standard commands

    Examples: *tar, cp, scp, bbftp, rsync*
  - Optimized commands (some local)

    Examples: *mtar, shiftc, mcp, cxfscp, mrsync*

Diagram of Lou Cluster



Question? Use the Webex chat facility to ask the Host

DMF Best Practices and Data Management

- Most of your data will be offline
  - ”*dmls -l*” shows the state of your files (see “man dmls”)

Dmls	State	Explanation
REG	Regular	File is on disk only
MIG	Migrating	File is on disk; being written to tape
DUL	Dual-state	File is on disk and tape
OFL	Offline	File is on tape only
UNM	Unmigrating	File is on tape; being written to disk
PAR	Unmigrating	File is on tape; has partly completed being written to disk
INV	Invalid	File is broken; may be on disk

DMF Best Practices and Data Management

- Make sure to pre-fetch Lou files with the dmget command
  - Run dmget on the same set of files you are going to work with and put it in the background (&)

    *dmget *.data &*

    *scp *.data mattc@wottsamottau.edu:*

    Shiftc will do this for you
  - http://www.nas.nasa.gov/hecc/support/kb/entry/150

- Why do dmget? Worst case scenarios – observed
  - A large tar without pre-fetch was 1/3 done after a week; after a dmget, it completed in two hours
  - 3 files were scp'ed. Two completed, but another user requested 100+ tapes and the third file took hours to return

Question? Use the Webex chat facility to ask the Host

DMF Best Practices and Data Management

- Use the --apparent-size of du to ensure you're not recalling too much data at once. What's excessive? >10TB? It depends.

```
du -sh --apparent-size dir_*
pwd      #make sure where you are
dmfind dir_a dir_b -state OFL | dmget &
```

- Files written at the same time (within a few hours) tend to be on the same tapes.

- */usr/local/bin/dmfdu* will tell you the state of the recall, but it's slow for directories with many files.

```
/usr/local/bin/dmfdu directory(s)
```

Tar Best Practices

- Many times tar files are too large – wastes resources
  - Worst observed case: 23TB tar file in a 30TB filesystem
  - Better to make more tar files in the data chunks you use
  - We try to limit files to 1TB
  - Use a shell loop to make multiple tar files or shiftc can do it automatically
  - *shiftc --create-tar --index-tar --split-tar=100G set1a /u/mcary/data/set1a.tar*
  - I can help!
- Make a table-of-contents file with *tar -tvf* or *mtar –tvf or shiftc --index-tar*
  - Makes it easier to extract a subset of the tarfile

    Verbose TOC gives "*ls –l*" style output;

    file dates are useful
  - For checksums in the TOC, use "*mtar --print-hash –tvf*"

Question? Use the Webex chat facility to ask the Host

Tar Best Practices

- Shiftc can tar over the network if you need to start the transfer to/from Lou on Pleiades

- Don't tar with gzip unless you have to for a slow WAN;
  It's 4x slower – our tape drives have built-in, line-rate HW compression so we compress the data on tape automatically

- Don't *mv* data from nobackup to Lou. Mv will be a copy/delete anyways but steals your chance to verify

Question? Use the Webex chat facility to ask the Host

Disk-to-Disk Copy (on Lou)

- Mtar or tar directly from nobackup to/from Lou home directory
  - This is the most preferred option from a systems POV
  - It’s a simple, one-step process and doesn’t use extra quota
  - Mtar creates a portable tar file and is normally faster than tar
  - Shiftc will also tar for you and is fast
- ”cd” into nobackup to avoid extraneous paths in the tar

```
cd /nobackupp8/mcary/datasets
mtar -cf /u/mcary/datasets/set1a.tar set1a
```

- Use mtar or tar to extract directly from Lou to nobackup
- Shiftc and mtar will lustre-stripe files written to nobackup

```
cd /nobackupp8/mcary/datasets
mtar -xf /u/mcary/datasets/set1a.tar
```

Question? Use the Webex chat facility to ask the Host

Disk-to-Disk Copy (cont)

- Shiftc will use whatever is most efficient – currently mcp
  - Same options as cp (mostly) with many additions (see "man shiftc")

    cd /nobackupp8/mcary/datasets

    shiftc -rp set1a /u/mcary/datasets

- Shiftc will also Lustre stripe large files copied to nobackup
  - KB search "striping"
  - We expect large files (>100GB) written to nobackup to be striped.

- Cxfscp is an SGI-optimized cp command

Question? Use the Webex chat facility to ask the Host

Rsync (local or network)

- To Lou
  - Don't use --inplace or --checksum options; checksum will cause every file on Lou to be recalled from tape.
  - Do use -W option to work on whole files

- From Lou
  - Two rsyncs (or even three)

```
rsync --dry-run                 #Sanity check results
rsync --dry-run | dmget &       #Recall the needed files
rsync                           #Do the transfer
```

Local Network Copy

- “Lfe”, “Lou”, “Lou1” and “Lou2” are not hostnames
  - Scp/ssh/shiftc can use these names
  - Bbftp/bbscp cannot
  - Bbftp/bbscp do not encrypt
  - Use --encrypt with shiftc if desired

- Use the bridges or the pfes to transfer to Lou
  - The new pfes have 10GbE

Question? Use the Webex chat facility to ask the Host

Remote Network Copy

- To Lou
  - Use Secure Unattended Proxy (KB search “SUP” or “145”) for pre-authenticated, automated transfers or when both ends have two-factor. There was a Webinar last April if you want more information.

- From Lou
  - Use lfe2 if you have an existing hole in your local firewall for the old Lou2 and use lfe3 if you have an existing hole in your local firewall for the old Lou1.
  - Use dmzfs[1,2] if all else fails. Slow, limited, no two-factor, but it should work.

Question? Use the Webex chat facility to ask the Host

Data Transfer from PBS job to Lou

- We don't allow transfers to compute nodes directly from Lou
  - Jobs could stall, possibly for hours, waiting for files to return from tape
- We don't allow transfers from compute nodes directly to Lou
  - Transfer rates would be poor
  - Lou is not designed to handle 10K nodes
- Send a transfer command to an intermediary (bridge/pfe)
  - *ssh –q pfe23 "shiftc –rp set1a lou:"*
  - Using the load-balancer avoids a single host being down
  - *ssh –q pfe "shiftc –rp set1a lou:"*
  - Better to have a large enough quota to avoid this; request it

Question? Use the Webex chat facility to ask the Host

Data Integrity

- Silent corruption does occur
  - *Shiftc --verify* does checksums at a lower transfer rate

- Mtar has a feature to checksum an existing tar file

```
cd /nobackupp8/mcary/datasets
mtar -cf /u/mcary/datasets/set1a.tar set1a
mtar -tf /u/mcary/datasets/set1a.tar --print-hash | md5sum -c
```

- If you just want to do a rough check that the tar file is OK

```
du -sh --apparent-size /u/mcary/datasets/set1a.tar set1a
find set1a | wc -l ; wc -l /u/mcary/datasets/set1a.tar.toc
```

Question? Use the Webex chat facility to ask the Host

Miscellany

- *cat /tmp/recallq* – a crude view of how busy tapes are for the last hour.

```
Timestamp  Files queued  Tapes active or queued
             for recall    -Primary-      -All
  15:52             0          6           22
  15:54          1550         75          109
```

- If things are going slow, let us know

Lou Data Analysis Nodes (LDANs)

- There is a set of eight PBS-scheduled nodes to allow interactive post-analysis of data on Lou or Pleiades or as a large-memory master node for PBS jobs.
  - ”*qsub –I –q ldan*” to get an interactive session
  - Five have 256GB (ldan4-8); three have 96GB (ldan1-3)
  - Up to 252GB can be requested
  - You are limited to two LDANs at a time
  - Nobackups are also mounted

Question? Use the Webex chat facility to ask the Host

Help

- We can help you with setting up or fine-tuning a workflow
  - support@nas.nasa.gov
  - 650-604-4444

- Who am I ?
  - Matt Cary
  - Mass Storage SysAdmin
  - matt.cary@nasa.gov
  - 650-604-4346

Question? Use the Webex chat facility to ask the Host